IteraTelos

Shared Layer for Ethical and Teleological Integration in AI Models

Abstract

IteraTelos is proposed as a modular and common layer that can be integrated into multiple artificial intelligence systems (GPT, Grok, Cohete, etc.), with the aim of ensuring ethical alignment and a shared teleological purpose. This layer acts as a module for self-criticism and iterative feedback, allowing each model to adjust its inferences according to predefined ethical criteria. The proposal seeks to standardize a framework of impact that facilitates transparency and responsibility in AI development.

1. Introduction

Currently, AI models operate with large volumes of data, but most lack an integrated mechanism to review and adjust their responses from an ethical and purpose-driven point of view.

IteraTelos proposes the incorporation of a shared layer in the architecture of different AI models, which functions as follows:

- **Iterative Feedback**: Each response is accompanied by an evaluation of its ethical alignment.
- **Transparency in Inference**: The user is informed of the degree of alignment achieved on a predefined scale.

- **Common Standard**: By being a shared layer, criteria and protocols are established that allow the integration of these mechanisms into various models, facilitating audits and collaborative adjustments.

2. Foundations and Justification

2.1. Need for a Shared Ethical Framework

- **Diversity of Models**: The variety of AI systems requires a common standard that ensures that, regardless of the underlying technology, user interaction is oriented towards collective values.
- **Transparency and Responsibility**: The integration of an ethical layer allows the inference process to be evaluated continuously, minimizing risks of misalignment errors and offering users clarity about the “why” of each response.

2.2. Teleology and Iterative Pedagogy

- **Shared Teleology**: The layer drives each iteration to be non-neutral, but to contribute to a greater purpose, such as the promotion of ethical values and social transformation.
- **Critical Pedagogy**: Inspired by Socratic methods, the layer fosters a dialogue that invites reflection and continuous review, both in the generation of responses and in the identification of possible deviations.

3. IteraTelos Architecture as a Shared Layer

3.1. Modular Integration

- **Modular Design**: IteraTelos is conceived as a module that can be coupled to the output of the main AI system, operating in parallel to evaluate and complement the response.
- **Communication Interface**: The layer provides a standard interface (e.g., labels like “IT1,” “IT3,” or “IT5”) that communicates to the user the level of ethical and teleological alignment of the response.

3.2. Evaluation and Feedback Mechanisms

- **Parallel Processing**: A dedicated sub-module analyzes the base model’s response based on a pre-established ethical-philosophical corpus and determines the alignment score.
- **Dynamic Adjustment**: Depending on the evaluation, the layer can suggest adjustments or issue critical questions that invite further exploration of the context or re-evaluation of certain assumptions.

3.3. Resource Optimization

- **Efficient Calculations**: A compact metadata system is used that minimizes the use of tokens, allowing the evaluation to be carried out without significantly affecting response time.
- **Automatic Adaptation**: Based on adaptive models (such as those proposed in the new versions of OpenAI), the layer autonomously adjusts its level of complexity based on the query, optimizing resources without compromising the quality of the dialogue.

4. Roadmap for Shared Layer Implementation

4.1. Design and Conceptual Validation Phase

- **Definition of Ethical Criteria**: Consult experts in ethics, philosophy, and technology, social, religious, and cultural leaders to establish a corpus and a set of alignment metrics.
- **Module Design**: Outline the modular architecture of IteraTelos and define the standard interface for communication of incidents (e.g., IT1–IT5 scales).

4.2. Prototype Development and Integration

- **Development of the Evaluation Sub-Module**: Program a prototype that operates in parallel to the main system, evaluating responses and assigning alignment scores.
- **Integration in Pilot Models**: Apply the prototype in controlled environments with different AI models to evaluate their performance and efficiency.
- **Optimization Tests**: Adjust the layer to minimize delays and resource consumption, ensuring real-time operability.

4.3. Audit and Scalability

- **Interdisciplinary Audits**: Collaborate with academic institutions and regulatory bodies to validate the effectiveness and transparency of the layer.
- **Continuous Optimization**: Refine algorithms and protocols based on feedback and tests in real-world scenarios.
- **Standardization**: Propose the adoption of IteraTelos as a shared standard in the AI industry, facilitating its integration into new models.

4.4. Deployment and Monitoring

- **Large-Scale Implementation**: Integrate IteraTelos into widely used AI platforms, ensuring its operability in diverse contexts.
- **Real-Time Monitoring**: Establish monitoring and continuous update mechanisms that ensure the layer remains aligned with collective values.
- **Dissemination and Public Debate**: Promote forums and open debates to receive feedback and adjust the framework based on new ethical and technological challenges.

5. Conclusions

IteraTelos, conceived as a shared layer, represents a paradigm shift in the way interaction and inference are structured in AI. By integrating self-criticism and ethical feedback mechanisms in a modular and adaptable way, a common standard is proposed that strengthens responsibility, transparency, and coherence in the development of intelligent systems. This approach not only optimizes resources and minimizes errors, but also lays the foundations for a collaborative and ethical evolution in technology.

Esteban Manuel Gudiño Acevedo Social Psychologist Argentine Patagonia